National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Data Fusion and Spatial Inference for Remote Sensing

Amy Braverman[1], Hai Nguyen[1], Emily Kang[3], Matthias Katzfuss[4], Pulong Ma[3], Anna Michalak[5], Noel Cressie[2], Tim Stough[1], and Vineet Yadav[1]

[1]Jet Propulsion Laboratory, Caltech
[2]NIASRA, University of Wollongong
[3]Department of Mathematical Sciences, University of Cincinnati
[4]Department of Statistics, Texas A&M University
[5]Carnegie Institution of Washington

June 12, 2017

Outline

- Introductory comments
- Carbon cycle science
- OCO-2 and AIRS data
- Exploiting synergy
- Example
- Data fusion strategy
- Spatial-statistical data fusion framework
- Spatial-statistical data fusion computation
- Spatio-temporal data fusion
- Validation example
- Take home
- Other applications and extensions
- References
- Acknowledgements

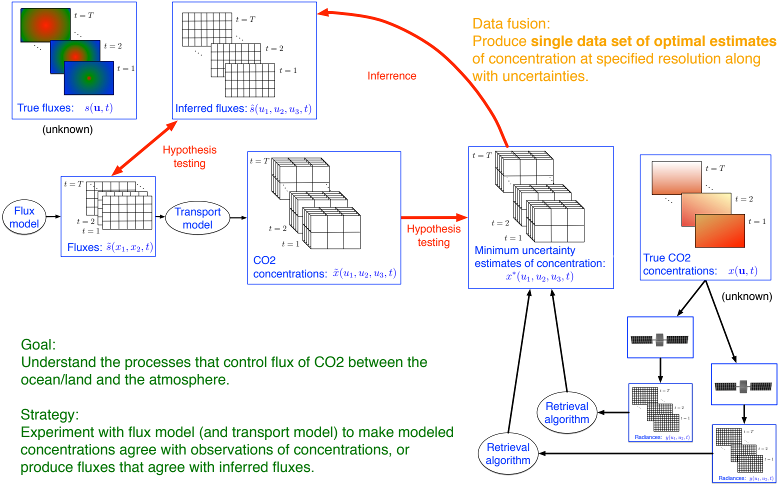
Introductory comments

- We want to estimate a complete geophysical field from massive, heterogeneous, observational data.
- The result is input to further science investigations and applications, so uncertainties must be propagated rigorously.
- Uncertainties should also be minimized so that conclusions, and decisions based on them, are as robust as possible. Need to leverage spatial and temporal dependencies.
- Challenge: Accomplish this in the face of massive data volumes and complex calculations required.

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Carbon cycle science



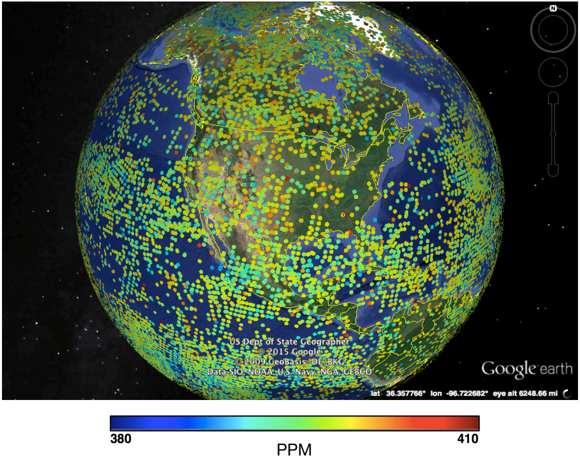
Goal:
Understand the processes that control flux of CO2 between the ocean/land and the atmosphere.

Strategy:
Experiment with flux model (and transport model) to make modeled concentrations agree with observations of concentrations, or produce fluxes that agree with inferred fluxes.

OCO-2 and AIRS data

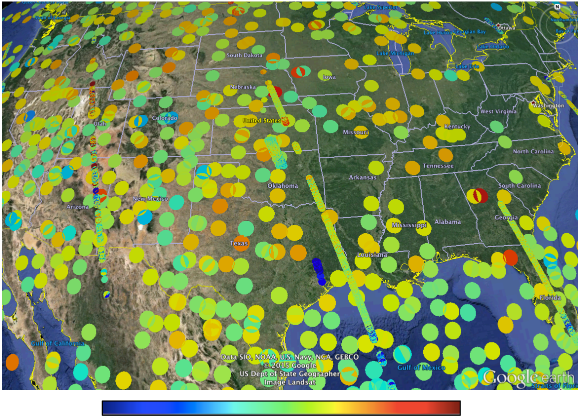
- OCO-2 and AIRS both observe column average CO2 mole-fraction, but are sensitive to different parts of the column.
- AIRS has a 90 km footprint, and OCO-2 has (about) a one km footprint.
- Their measurement errors and patterns of missingness are also different because they exploit different technologies and retrieval algorithms.

OCO-2 and AIRS data



AIRS mid-tropospheric CO_2, October 30 through November 2, 2014.

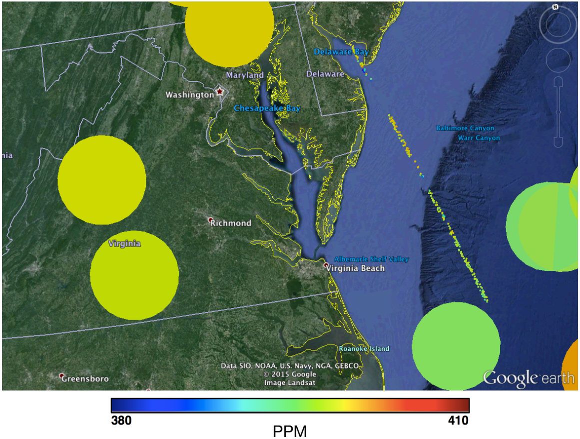
OCO-2 and AIRS data



AIRS mid-tropospheric and OCO-2 total column CO2, October 30 through November 2, 2014.

OCO-2 footprint size x10 for visualization.

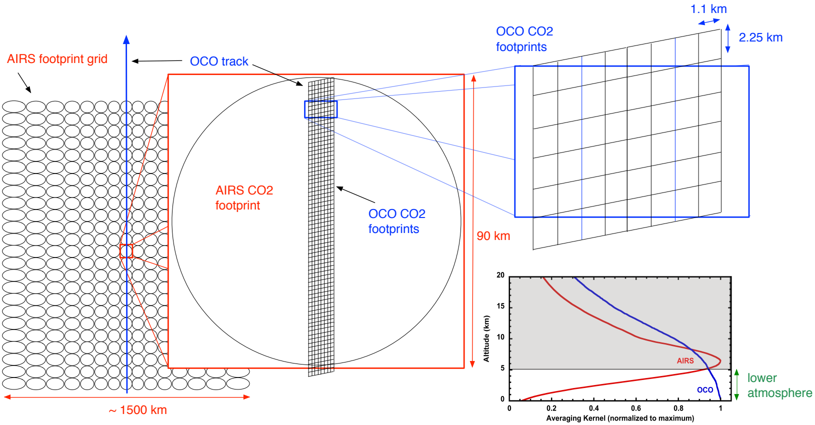
OCO-2 and AIRS data



AIRS mid-tropospheric and OCO-2 total column CO2, October 30 through November 2, 2014.

OCO-2 footprints actual size.

OCO-2 and AIRS data



Exploiting synergy

- Instrument sensitivities are similar at and above the mid-troposphere, but not below: OCO-2 is sensitive down to the surface, but AIRS is not.

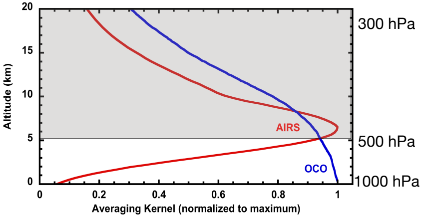
- To the extent that CO2 mole-fraction near the surface and in the mid-troposphere are correlated, we should be able to improve estimates of both by exploiting this correlation.

- We should also be able to
  - exploit the coverage of AIRS and the accuracy of OCO-2
  - exploit spatial and temporal correlations within and between data sets.

Exploiting synergy

Can these data be combined to create a more complete data set with information about CO2 closer to the surface?

Exploiting synergy



- If we knew the “true" values of total-column and mid-tropospheric mole-fraction at a location $\mathbf{s}$ =lat,lon, $Y_1(\mathbf{s})$ and $Y_2(\mathbf{s})$, then we could compute

$$Y_{LA}(\mathbf{s}) = \frac{(1000 - 300)\,Y_1(\mathbf{s}) - (500 - 300)\,Y_2(\mathbf{s})}{1000 - 500}.$$

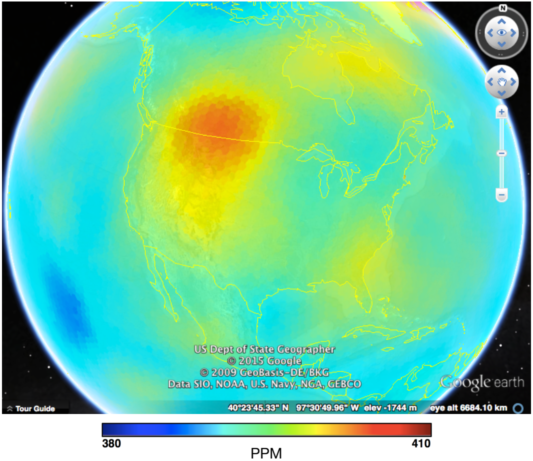
- Can we get estimates, with uncertainties, of (total-column, mid-trop) pairs at reasonable resolution so we can compute this?

Example

- Accumulate 12 days of AIRS and OCO-2 data into three, four-day blocks: Oct 30 - Nov 2, Nov 3 - 6, Nov 7 - 10.
- Run Spatio-Temporal Data Fusion algorithm (STDF) on the three blocks, producing three output data sets, one for each block. (See Nguyen, Katzfuss, Cressie, and Braverman (2014) for details.)
- STDF accounts for spatial correlations among footprints for both instruments (including corrections for different sizes and orientations) and for temporal correlations from time block to time block.
- Timing: 90 minutes to process the three blocks on a single, Intel Xeon 2.0 GHz processor.
- Crucial assumptions: uncertainty on AIRS datum is 1.5 ppm, and uncertainty on OCO-2 datum is 2 ppm.

Example

Fused estimate of lower-atmosphere CO2, Oct 30 - Nov 2, 2014:



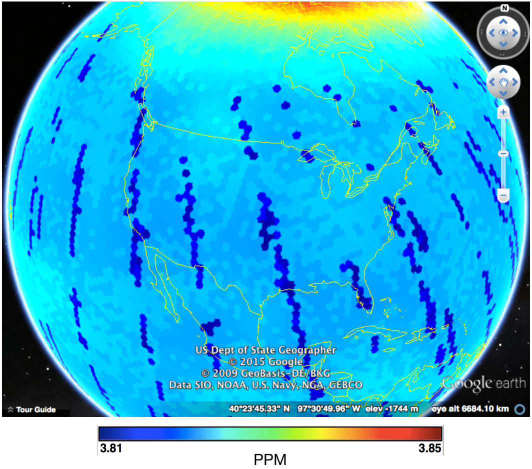
Produced using STDF with analysis resolution ≈ 30 km.

Visualization resolution ≈ 120 km.

How to validate estimates?

Example

Uncertainties of fused uncertainties, Oct 30 - Nov 2, 2014:



Lower uncertainties coincide with OCO-2 tracks.

How to validate uncertainties?

Example

Caveats:

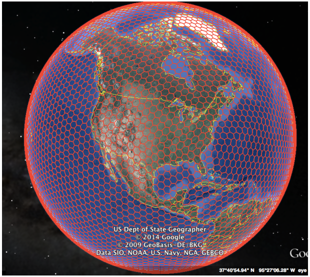
- OCO-2 data are very preliminary: just a placeholder here to show data fusion machinery.
- The formula for computing lower-atmosphere mole-fraction is unrealistically crude.
- Uncertainties on the input data are unrealistic (but the best we've got right now). This is a *major* issue.
- We have built a simulation system for characterizing the performance of STDF on synthetic “truth" data, and are in the process of assessing how various design choices affect our results.

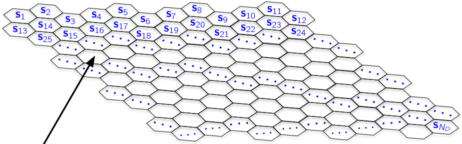
National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Data fusion strategy

- In order to do this calculation, we need to infer the true mole-fractions of (total-column, mid-trop) pairs on a fine grid of locations.
- We define that grid by partitioning the world into very small hexagonal tiles called basic areal units (BAU's) Notionally, each BAU contains a pair.



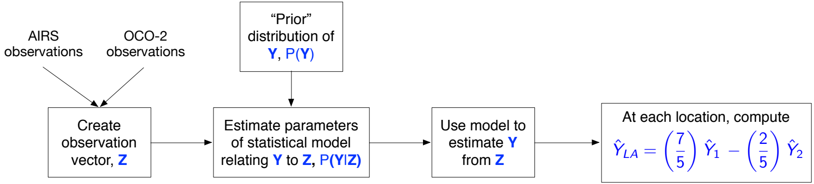


$$Y(\mathbf{s}) = (Y_1(\mathbf{s}), Y_2(\mathbf{s}))$$

Data fusion strategy

- Since this bivariate field is unknown, we model it with a random vector that behaves according to a probability distribution.
- We use Bayes' Theorem: before acknowledging the observations, we assume a "prior" distribution.
- After seeing the data, we update that distribution and call it the "posterior".
- We report the mean vector and covariance matrix of the posterior distribution as our inference.

Spatial-statistical data fusion framework



Spatial-statistical data fusion framework

AIRS observations

OCO-2 observations

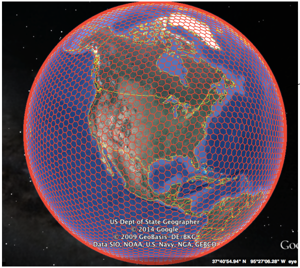
"Prior" distribution of **Y**, P(**Y**)

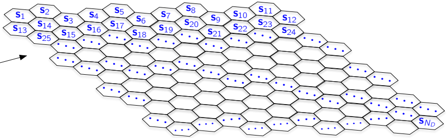
Create observation vector, **Z**

Estimate parameters of statistical model relating **Y** to **Z**, P(**Y**|**Z**)

Use model to estimate **Y** from **Z**

At each location, compute $\hat{Y}_{LA} = \left(\frac{7}{5}\right)\hat{Y}_1 - \left(\frac{2}{5}\right)\hat{Y}_2$

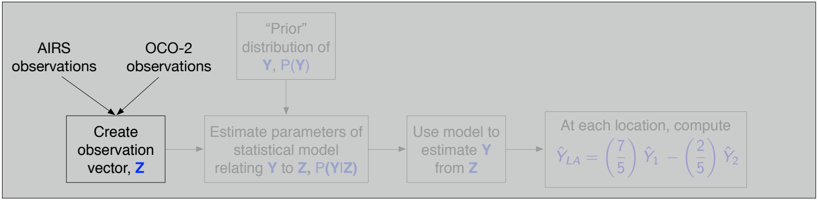


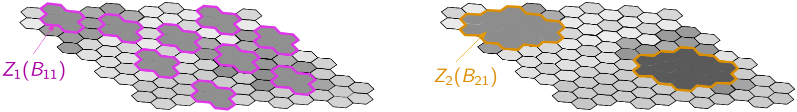


$$\mathbf{Y} = [Y_1(\mathbf{s}_1), \ldots, Y_1(\mathbf{s}_{N_D}), Y_2(\mathbf{s}_1), \ldots, Y_2(\mathbf{s}_{N_D})]$$

$$\mathbf{Y} \sim N(\boldsymbol{\mu}_\mathbf{Y}, \boldsymbol{\Sigma}_\mathbf{Y})$$

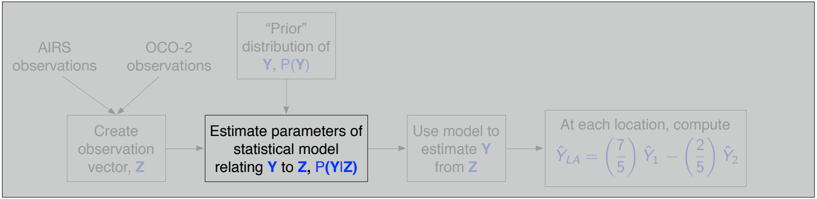
Spatial-statistical data fusion framework

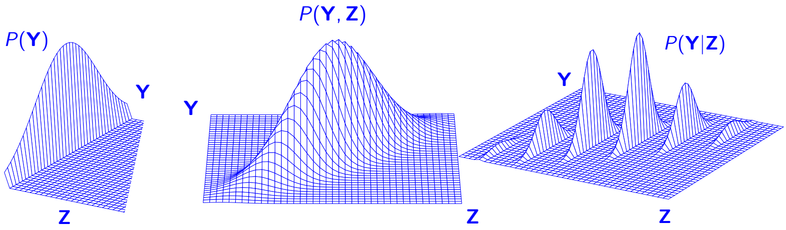




$$\mathbf{Z} = [Z_1(B_{11}), \ldots, Z_1(B_{1N_1}), Z_2(B_{21}), \ldots, Z_2(B_{2N_2})]$$

Spatial-statistical data fusion framework





Spatial-statistical data fusion framework

AIRS observations

OCO-2 observations

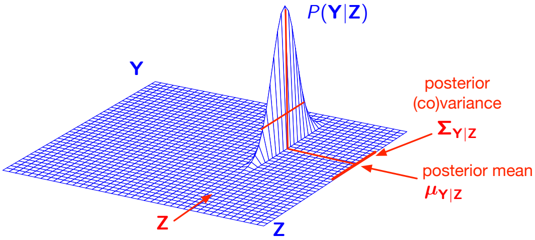
"Prior" distribution of **Y**, P(**Y**)

Create observation vector, **Z**

Estimate parameters of statistical model relating **Y** to **Z**, P(**Y**|**Z**)

Use model to estimate **Y** from **Z**

At each location, compute $\hat{Y}_{LA} = \left(\frac{7}{5}\right)\hat{Y}_1 - \left(\frac{2}{5}\right)\hat{Y}_2$



Spatial-statistical data fusion framework

AIRS observations

OCO-2 observations

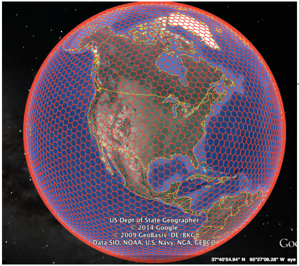
"Prior" distribution of **Y**, P(**Y**)

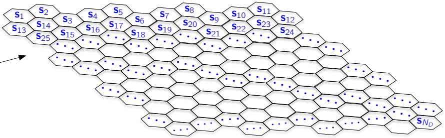
Create observation vector, **Z**

Estimate parameters of statistical model relating **Y** to **Z**, P(**Y**|**Z**)

Use model to estimate **Y** from **Z**

At each location, compute $\hat{Y}_{LA} = \left(\frac{7}{5}\right)\hat{Y}_1 - \left(\frac{2}{5}\right)\hat{Y}_2$





$$\mathbf{Y} = [Y_1(\mathbf{s}_1), \ldots, Y_1(\mathbf{s}_{N_D}), Y_2(\mathbf{s}_1), \ldots, Y_2(\mathbf{s}_{N_D})]$$

$$\mathbf{Y} \sim N(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{Z}}, \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{Z}})$$

Spatial-statistical data fusion computation

- At 30 km analysis resolution there are 660,000 BAU's over the globe. (At 1 km, there are about 700,000,000.)

- Over a four day time block (the temporal snapshot we use), there are about 180,000 observations total from both instruments.

- The formulas for the posterior mean and covariance of the field given the observations can't be implemented as-is: the problem is too big.

Spatial-statistical data fusion computation

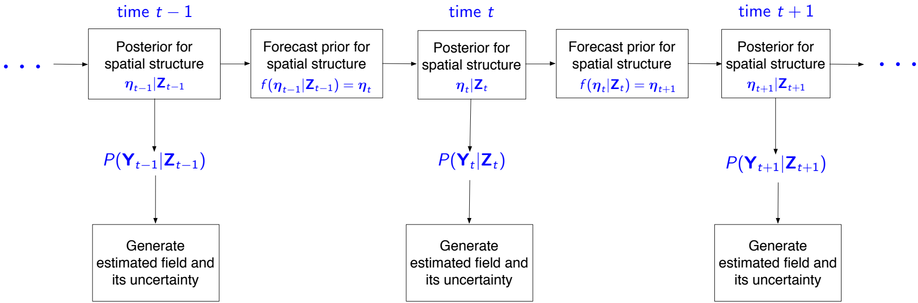
- Impose some additional constraints and modeling assumptions on $\mathbf{Y}$.

- Key: spatial relationships in the field $\mathbf{Y}$ admit a simpler, low-dimensional representation in the form of a hidden spatial structure variable, η, defined relative to a set of fixed spatial basis functions.

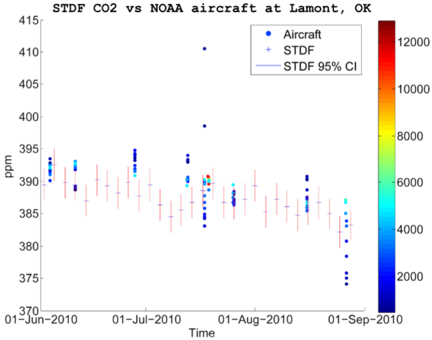
- Posterior distribution of $\mathbf{Y}$ given $\mathbf{Z}$ is found by first obtaining an estimate of the posterior distribution of η given $\mathbf{Z}$, then reconstructing the posterior distribution of $\mathbf{Y}$ from it.

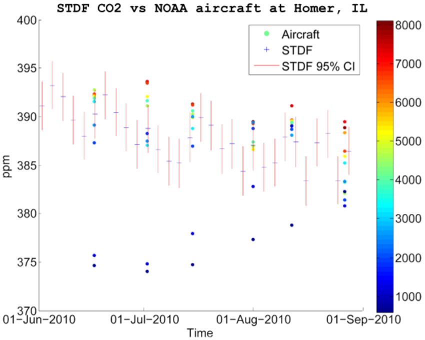
Spatio-temporal data fusion

- Temporal dependence from time block to time block is exploited by Kalman filtering (or, in our case, smoothing) $\boldsymbol{\eta}$.



Validation example





- Validation of STDF estimates of lower-atmosphere CO2 based on AIRS and Japan’s Greenhouse Gases Observing Satellite (NASA retrievals). See Nguyen et al. (2014) for details.

Take home

- Data fusion is necessary to realize benefits of synergy among NASA missions.

- What is new about this data fusion technology:
  - based on uncertainty quantification and minimization
  - uses a formal probabilistic framework that is coherent
  - exploits spatial and temporal correlations to drive uncertainties down
  - corrects for heterogeneous footprints
  - feasible for massive data sets and operational implementation.

- Better results are possible if mission provide formal uncertainty estimates for their retrieivals.

Other applications and extensions

Other possible applications (infusion):

- Aerosol optical depth from MISR and MODIS-Terra (case study in Nguyen, Cressie, and Braverman, 2012).
- Sea-surface temperature from MODIS-Terra, MODIS-Aqua, VIIRS, and AMSR-2.
- Surface temperature from AIRS, CrIS (and IASI?).
- OCO-2 CO2 and fluoresence, SMAP soil moisture, and MODIS fraction photosynthetically active radiation, and leaf-are index (possible future).

Other applications and extensions

Extensions:

- Fusion of multivariate quantities, e.g, atmospheric profiles. See Nguyen, Cressie, and Braverman (2017).
- Adaptive grids: high-resolution in region of interest, lower resolution elsewhere.
- Data fusion in distributed environments: data fusion without moving data.

References

The spatio-temporal data fusion methodology used here is described in detail in

Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2014). Spatio-Temporal Data Fusion for Very Large Remote Sensing Datasets, *Technometrics*, 56, pp. 174-185.

The spatial-only methodology is described in

Nguyen, H., Cressie, N., and Braverman, A. (2012). Spatial Statistical Data Fusion for Remote-Sensing Applications, *Journal of the American Statistical Association*, 107, pp. 1004-1018.

The extension to the fusion of profiles is described in

Nguyen, H., Cressie, N., and Braverman, A. (2017). Multivariate Spatial Data Fusion for Very Large Remote Sensing Datasets, *Remote Sensing*, 9(2), pp. 1004-1018, DOI:10.3390/rs9020142.

This work was partially performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Government sponsorship acknowledged.

This research was funded by NASA's Earth Science Technology Office through its Advanced Information Systems Technology program (AIST-08 and AIST-11).

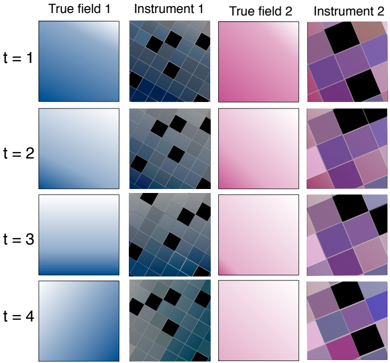
We would like to thank the OCO-2 and AIRS Projects for their intellectual and practical contributions to this work.

Contact information: Amy.Braverman@jpl.nasa.gov

Backup slides

Spatial inference and data fusion

Exploit spatial correlations	Exploit spatial and temporal correlations
Infer the true field (single quantity) from one remote sensing image of it at a single time point. (Fixed Rank kriging)	Infer the true field (single quantity) from one remote sensing image of it at multiple time points. (Fixed Rank filtering or smoothing)
Infer the true field from *two* different remote sensing images of it at a single time. (Single process, multiple source spatial data fusion)	Infer the true field from *two* different remote sensing images of it at multiple time points. (Single process, multiple source spatio-temporal data fusion)
Infer true values of *two* fields from *two* different remote sensing images at a single time. (Muliple process, multiple source spatial data fusion)	Infer true values of *two* fields from *two* different remote sensing images at multiple time points. (Muliple process, multiple source spatio-temporal data fusion)



Bayes' Theorem

- Given two events, A and B,

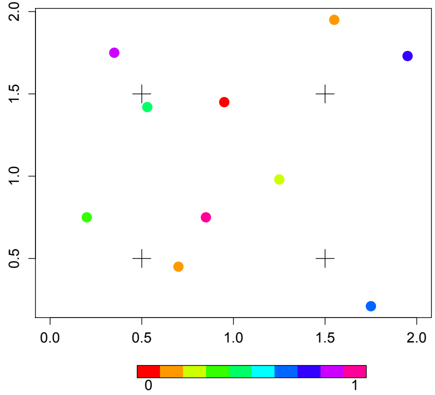
$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}.$$

- Example: B = event that the freeway is jammed, A = event the on-ramp is jammed.

$$P(\text{freeway jammed}|\text{on-ramp jammed}) = \frac{P(A|B)P(B)}{P(A)},$$
$$= \frac{P(\text{on-ramp jammed}|\text{freeway jammed})P(\text{freeway jammed})}{P(\text{on-ramp jammed})}.$$

Spatial dimension reduction

$+ =$ basis center,
$\bullet =$ spatial location, color $=$ value



$$\boldsymbol{\nu} = (\nu(\mathbf{s}_1), \ldots, \nu(\mathbf{s}_{10}))'$$

- Spatial field: values at ten locations, $\boldsymbol{\nu}$.
- Spatial structure described by spatial covariance matrix, $\boldsymbol{\Sigma}_{\boldsymbol{\nu}}$ (10×10).
- Basis centers are reference locations.

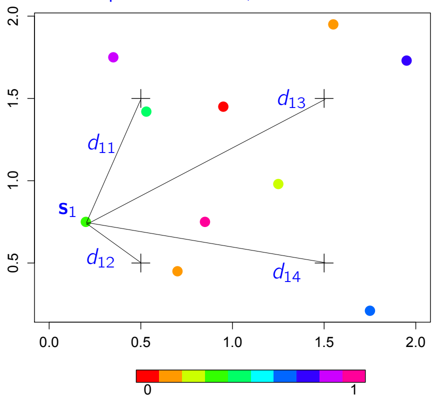
National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Spatial dimension reduction

$+ =$ basis center,

$\bullet =$ spatial location, color $=$ value



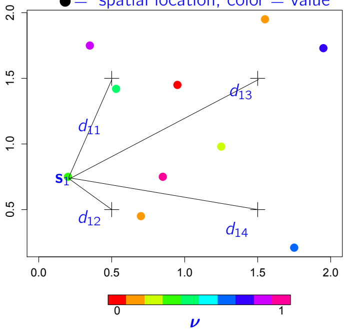
- Spatial field: values at ten locations, $\boldsymbol{\nu}$.
- Spatial structure described by spatial covariance matrix, $\boldsymbol{\Sigma}_{\nu}$ (10×10).
- Basis centers are reference locations.
- Encode each location as inverse of distances to four basis centers:

$$\mathbf{S}(\mathbf{s}_1) = (1/d_{11}, 1/d_{12}, 1/d_{13}, 1/d_{14}).$$

Spatial dimension reduction

$+ =$ basis center,

$\bullet =$ spatial location, color $=$ value



Basis function matrix:

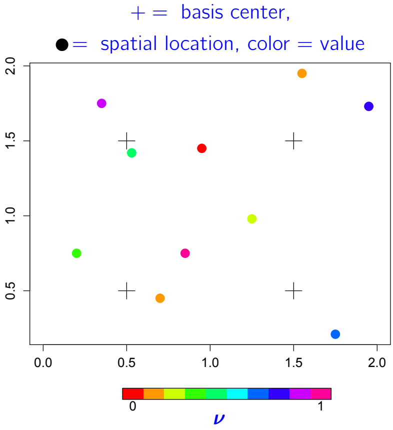
$$\mathbf{S} = \begin{pmatrix} \mathbf{S}(\mathbf{s}_1) \\ \mathbf{S}(\mathbf{s}_2) \\ \vdots \\ \mathbf{S}(\mathbf{s}_{10}) \end{pmatrix} = \begin{pmatrix} 1/d_{11} & 1/d_{12} & 1/d_{13} & 1/d_{14} \\ 1/d_{21} & 1/d_{22} & 1/d_{23} & 1/d_{24} \\ \vdots & \vdots & \vdots & \vdots \\ 1/d_{10,1} & 1/d_{10,2} & 1/d_{10,3} & 1/d_{10,4} \end{pmatrix}$$

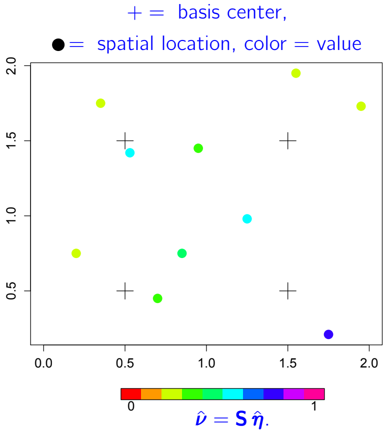
Low-dimensional representation:

$$\boldsymbol{\nu} = \mathbf{S}\,\boldsymbol{\eta} = \begin{pmatrix} 1/d_{11} & 1/d_{12} & 1/d_{13} & 1/d_{14} \\ 1/d_{21} & 1/d_{22} & 1/d_{23} & 1/d_{24} \\ \vdots & \vdots & \vdots & \vdots \\ 1/d_{10,1} & 1/d_{10,2} & 1/d_{10,3} & 1/d_{10,4} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{pmatrix}$$

- Estimate $\boldsymbol{\eta}$ by least-squares (for example).
- Linearity: $\boldsymbol{\nu} = \mathbf{S}\boldsymbol{\eta} \Longrightarrow \boldsymbol{\Sigma}_{\nu} = \mathbf{S}\boldsymbol{\Sigma}_{\eta}\mathbf{S}'$. $\boldsymbol{\Sigma}_{\eta}$ is only 4×4.

Spatial dimension reduction





- Reconstructed field is an approximation to the original, but much more parsimonious.